文章编号:1671 - 6833(2017)06 - 0029 - 04

基于混合差分演化的网络入侵检测算法

王耀光[1], 陈伟权[1], 吴镇邦[1], 秦 勇[2], 黄 翰[3]

(1.广东省东莞市质量监督检测中心,广东 东莞 523000; 2.东莞理工学院 计算机学院, 广东 东莞 523000; 3.华南理工大学 软件学院,广东 广州 510006)

摘 要: 基于机器学习方法的入侵检测算法是目前网络设备检测领域的研究热点.网络入侵检测源数据的多样性是影响机器学习方法在该领域实际应用性能的主要因素.研究通过设计多扰动向量混合差分演化算法,稳定地优化了最小二乘支持向量机模型的关键参数;在不增加测试集检测计算复杂性的前提下,通过最优化参数的方式,提高了最小二乘支持向量机算法入侵检测的精度和稳定性.KDD Cup 99 测试集的仿真实验结果显示,所提出的基于混合差分演化的网络入侵检测算法比目前多种同类算法有着更好的平均性能.

关键词: 网络入侵检测; 测试稳定性; 混合差分演化; 最小二乘支持向量机

中图分类号: TP301.6 **文献标志码**: A doi:10.13705/j.issn.1671 - 6833.2017.06.006

0 引言

入侵检测系统 IDS(intrusion detection system)是计算机信息安全领域的一个重要分支.IDS 的特点是主动防御,即对入侵行为进行预警,关键技术是对入侵事件的识别和分类上.近年来,机器学习与智能计算相结合的方法成为了网络入侵检测研究的热门技术.Hu 等[1]利用粒子群算法对在线 Adaboost 的参数进行优化,解决了动态分布式网络入侵问题.刘羿[2]提出了蝙蝠算法优化神经网络,提高了网络入侵检测的效率.李振刚等[3]运用改进的蚁群算法来优化 SVM 参数,提升了入侵检测的效率和准确率.王亚等[4]研究发现通过优化 RBF 神经网络的参数,可以有效降低特征维数和 RBF 神经网络输入节点数,从而降低计算复杂度,加快网络入侵检测速度.Dastanpour 等[5]系统研究了遗传算法在在优化神经网络和支持向量机参数上的效果,发现网络入侵检测源数据的多样性是影响演化算法实际应用性能的主要因素.实际工程应用对网络入侵检测的准确率和响应速度要求较高,因此,高速、高准确率的要求是困扰智能计算方法应用于网络入侵检测的核心难题.

为了解决这一难题,差分演化算法和网络入侵检测的结合成为了研究热点.马琰等[6]将混沌差分算法用于网络入侵检测,降低了检测的误差;边根庆等[7]将免疫克隆与差分进化相统一,为进化算法在网络入侵检测中的实际应用做了理论铺垫;Sailaja 等[8]基于差分演化算法对网络入侵检测问题进行了研究.考虑到网络入侵检测源数据的多样性,笔者采用了多扰动向量的混合差分演化算法,对最小支持向量机(LSSVM)[3]的大量关键参数进行优化,通过对于差分演化算法的改进,使得优化后的 LSSVM 提高了网络入侵检测的准确率和稳定性.

1 基于最优参数 LSSVM 网络入侵检测的可行性分析

由于入侵事件种类的多样性和稳定性,网络入侵检测可以建模为一个模式分类问题.许多网络入侵检测技术都应用最小支持向量机(LSSVM)来进行入侵事件的分类;然而,LSSVM 的分类准确率受其多个参数的影响.通过参数优化来提高网络入侵检测精度的研究是目前学术界研究的热点,其技术路线如图 1 所示.

LSSVM 算法定义如公式(1):

收稿日期:2017 - 05 - 20;**修订日期**:2017 - 07 - 18

基金项目:国家自然科学基金资助项目(61370102);广东省高等院校学科与专业建设专项资金建设项目(2013KJCX0178)

作者简介:王耀光(1964—),男,广东东莞人,广东省东莞市质量监督检测中心高级工程师,主要从事信息技术设备安全质量评估方面的研究,E-mail:wyg@gddqt.com.

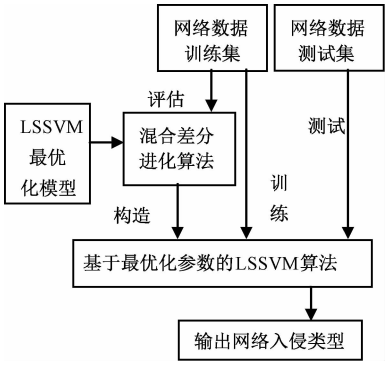

图 1 基于 LSSVM 模式分类的网络入侵检测技术路线图

Fig. 1 Technology roadmap of network intrusion detection based on LSSVM model classification

$$\begin{cases}\min\ J(\boldsymbol{\omega},\xi) = \dfrac{1}{2}(\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{n}\xi_i^2);\\ \text{s.t. } y_i[\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i) + b] - 1 + \xi_i = 0.\end{cases} \tag{1}$$

可以通过引入拉格朗日乘子将式(1)的模型变化为式(2)的实优化问题:

$$\min L = J(\boldsymbol{\omega},\xi) - \sum_{i=1}^{n} a_i\{y_i\cdot[\boldsymbol{\omega}^{\mathrm{T}}\varphi(x_i) + b] - 1 + \xi_i\}. \tag{2}$$

一般的做法是将式(2)对 $\boldsymbol{\omega}$、ξ_i、b 和 a_i 求偏导,通过消去 $\boldsymbol{\omega}$ 和 ξ_i 得到关于 b 和 a_i 的方程组,用数值计算的方法可以求解方程组并得到最优化的 b 和 a_i,$i = 1,2,\cdots,n$,n 为采样的规模. 因为数值计算的方法所需计算时间长且复杂性较大,一些 LSSVM 应用为了节省计算时间而采用了近似算法,如粒子群优化算法等[3]. 但是,因为网络入侵检测的实时性要求较高,故而采用更高效的连续优化启发式算法来求得最优的 b 和 a_i.

在实际工程应用中可以运用训练集样本来评估 LSSVM 优化的效果. 如果用于评估的样本足够描述网络入侵数据的特征,那么优化后的 LSSVM 模型可以高精确度地分类出网络入侵事件. 由于最优参数的 LSSVM 算法在实际检测过程中不再需要增加额外的计算量,因此,图 1 所示的方法在计算时间复杂性上是可行的.

2 改进算法设计

考虑到网络入侵检测的实时性与精确性要求,因此采用了差分演化算法(DE)[9]作为优化 LSSVM 参数的核心技术,而 DE 是目前解决连续优化问题最有效的算法之一. 虽然 DE 算法在单目标连续优化问题的求解上表现出了较强的性能,但是在求解式(2)的优化问题时容易陷入局部最优解和收敛慢的困境,而扰动向量是对 DE 算法性能影响最大的一个因素. 因此,笔者通过采用多个扰动向量设计扰动向量池来实现混合差分变异规则,从而提高算法的全局搜索能力和收敛速度.

2.1 基于扩展扰动向量池的混合差分演化算法

差分演化算法的目标是给不断演化的 N 个 D 维的向量找到全局最优,式(2)的 LSSVM 优化参数 b 和 a_i 可编码为 $\boldsymbol{X}_i^G = \{x_{i,1}^G, x_{i,2}^G, \cdots, x_{i,D}^G\}$,$i = 1,2,\cdots,N$. 其中,$G$ 为当前种群迭代次数; D 为目标函数维度大小. 初始的向量分布最好均匀地分布在整个搜索空间内,搜索空间的最大和最小边界被预先设定:$\boldsymbol{X}_{\min} = \{x_{\min,1}, x_{\min,2}, \cdots, x_{\min,D}\}$ 和 $\boldsymbol{X}_{\max} = \{x_{\max,1}, x_{\max,2}, \cdots, x_{\max,D}\}$. 借助文献[4]的方法可以将网络数据的字符映射为数值,$\boldsymbol{X}_{\min}$ 和 $\boldsymbol{X}_{\max}$ 根据这些数值的范围确定.

考虑到网络数据的多样性与入侵事件类型的稳定性,笔者设计了混合策略用来改进差分演化算法的性能. 在混合策略中,将针对每一个混合步骤个体的变异设计多个扰动向量来生成新个体进入下一代,通过混合策略可以计算多个扰动向量的新子代,然后选取最优的进入下一代. 混合策略可以被分为两类:①计算每一个扰动向量产生子代的最优值,如果其中有优过父代的则选取最优的一个进入下一代,反之继续沿用父代个体;②构建一个选择模型用来预测每一代应该通过哪一个扰动向量来产生新个体. 通常一个好的预测模型可以在大多数迭代过程中选择最优的扰动向量. 候选扰动向量的列表如式(3)~(15)所示.

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{\text{best}}^G + \boldsymbol{F}(\boldsymbol{X}_{r1}^G - \boldsymbol{X}_{r2}^G); \tag{3}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{\text{best}}^G + \boldsymbol{F}(\boldsymbol{X}_{r1}^G - \boldsymbol{X}_{r2}^G) + \boldsymbol{F}(\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r4}^G); \tag{4}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}\cdot(\boldsymbol{X}_{r2}^G - \boldsymbol{X}_{r3}^G); \tag{5}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}(\boldsymbol{X}_{r2}^G - \boldsymbol{X}_{r3}^G) + \boldsymbol{F}(\boldsymbol{X}_{r4}^G - \boldsymbol{X}_{r5}^G); \tag{6}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_i^G + \boldsymbol{F}(\boldsymbol{X}_{\text{best}}^G - \boldsymbol{X}_i^G) + \boldsymbol{F}(\boldsymbol{X}_{r1}^G - \boldsymbol{X}_{r2}^G); \tag{7}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}(\boldsymbol{X}_{\text{best}}^G - \boldsymbol{X}_{r2}^G) + \boldsymbol{F}(\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r4}^G), \tag{8}$$

式中:$\boldsymbol{F}(\boldsymbol{X}_{r2}^G - \boldsymbol{X}_{r3}^G)$ 是一个权重的差分向量算子; $\boldsymbol{X}_{\text{best}}^G$ 是第 G 代的最优个体;$\boldsymbol{F}$ 是一个大于零的控制向量. 式(3)~(8)的扰动向量设计参考了 Storn 和 Price 提出的差分演化算法扰动向量家族[9].

由于公式(2)的优化模型由网络数据的样本向量决定,基于网络数据的多样性,图 1 所示的技术路线将造成优化模型的动态变化,因此,单一的优化策略难以有效地解决该问题. 与经典 LSSVM

参数优化算法[3-4]不同,混合使用多种扰动向量可以增加差分演化算法种群的多样性.因此,可以参考基因算法的情况,每个个体可以选取同样的个体来产生新的子代,这在数学上增加了个体的多样性,参见图 2.

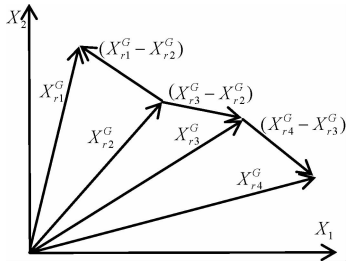

图 2　在 2 维空间内解释 DE 扰动向量组件

Fig. 2　Explained disturbance vector components of DE in the two-dimensional space

新的扰动向量生成策略增加了种群多样性的潜力,($\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r2}^G$) 就是增加的差分向量的选项.针对网络样本数据的多样性,笔者扩展设计了新的扰动向量,如式(9)至(15)所示:

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{\text{best}}^G + \boldsymbol{F}(\boldsymbol{X}_{r1}^G - \boldsymbol{X}_{r2}^G) + \boldsymbol{F}(\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r4}^G) + \boldsymbol{F}(\boldsymbol{X}_{r5}^G - \boldsymbol{X}_{r6}^G); \tag{9}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}(\boldsymbol{X}_{r2}^G - \boldsymbol{X}_{r3}^G) + \boldsymbol{F}(\boldsymbol{X}_{r4}^G - \boldsymbol{X}_{r5}^G) + \boldsymbol{F}(\boldsymbol{X}_{r6}^G - \boldsymbol{X}_{r7}^G); \tag{10}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{\text{best}}^G + \boldsymbol{F}(\boldsymbol{X}_i^G - \boldsymbol{X}_{r1}^G) + \boldsymbol{F}(\boldsymbol{X}_{r2}^G - \boldsymbol{X}_{r3}^G); \tag{11}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_i^G + \boldsymbol{F}(\boldsymbol{X}_{r1}^G - \boldsymbol{X}_{r2}^G) + \boldsymbol{F}(\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r4}^G); \tag{12}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}(\boldsymbol{X}_i^G - \boldsymbol{X}_{r2}^G) + \boldsymbol{F}(\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r4}^G); \tag{13}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}(\boldsymbol{X}_{\text{best}}^G - \boldsymbol{X}_{r1}^G) + \boldsymbol{F}(\boldsymbol{X}_{r2}^G - \boldsymbol{X}_{r3}^G) + \boldsymbol{F}(\boldsymbol{X}_{r4}^G - \boldsymbol{X}_{r5}^G); \tag{14}$$

$$\boldsymbol{V}_i^G = \boldsymbol{X}_{r1}^G + \boldsymbol{F}(\boldsymbol{X}_{\text{best}}^G - \boldsymbol{X}_{r2}^G) + \boldsymbol{F}(\boldsymbol{X}_{r3}^G - \boldsymbol{X}_{r4}^G) + \boldsymbol{F}(\boldsymbol{X}_{r5}^G - \boldsymbol{X}_{r6}^G). \tag{15}$$

式(9)~(15)提出的扰动向量兼顾了局部搜索以及全局搜索的能力,在多样性的优化问题上比较容易取得稳定的效果,可以弥补 Storn 和 Price 提出的差分演化扰动向量家族的不足.

2.2　改进差分演化算法的流程

按照式(3)~(15)选取扰动向量,流程如图 3 所示.启发式步骤包括:在差分演化算法初始化时,随机从扰动向量池内选择一个扰动向量;运行差分演化算法时,如果所选的扰动向量没有进一步优化公式(2)的问题,重新随机选择一个不同于之前所选的扰动向量.

当基于新网络数据的优化问题生成时,即某个扰动向量不能使差分演化算法得到更优解时,可以重新选择扰动向量池内的扰动向量.而扰动向量可以根据不同的优化模型来选取,式(3)~(15)提供了备选扰动向量更新公式,它们都具有良好的可装卸性.通过拆卸和组装混合扰动向量池的扰动向量组合,可以更有效方便地解决不同类型网络数据在图 1 技术路线下形成的优化问题:如单峰问题适合使用 best 驱动的扰动向量;多峰问题适合利用 rand 偏移来跳出局部最优解.

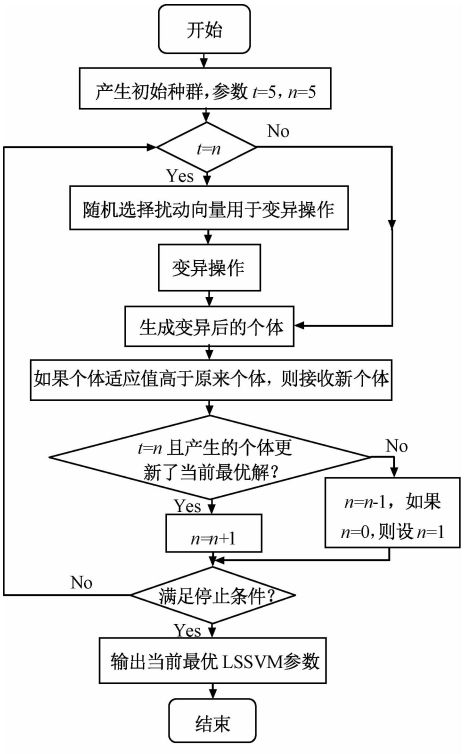

图 3　改进差分演化算法的流程

Fig. 3　The process of improved differential evolution algorithm

3　仿真实验

3.1　实验设置

本实验采用 KDD CUP 99 数据进行离线测试,验证基于混合差分演化的网络入侵检测算法(简记为 HDE-SVM)的效率.笔者提出的 HDE-LSSVM 算法包含了差分演化和最小支持向量机两部分,参数设置如下:LSSVM 公式(2)中的参数由 HDE 算法优化确定,采样规模 n 设置为 30.根据文献[9-10]建议,HDE 种群规模 N 设置为 100,惯性权重 $w = 0.5$,$CR = 0.9$.

3.2　实验结果与分析

笔者提出的 HDE-SVM 算法采用 6 个节点进行测试,通过 KDD CUP 99 的数据测试对比了 PSO-SVM、PSO-LSSVM、投票方法[11]、DE-LSSVM、SaDE-LSSVM[12]和 HDE-LSSVM 的性能,研究结果以检测准确率和误报率作为对比指标,实验结果如表 1 所示.

表1 PSO-SVM、PSO-LSSVM、投票方法、DE-LSSVM、SaDE-LSSVM 和 HDE-LSSVM 的对比实验结果

Tab.1 The contrast experiment results of PSO-SVM, PSO-LSSVM, voting method, DE-LSSVM, SaDE-LSSVM and HDE-LSSVM %

节点	PSO-SVM		PSO-LSSVM		投票方法		DE-LSSVM		SaDE-LSSVM		HDE-LSSVM	
	准确率	误报率	准确率	误报率	准确率	误报率	准确率	误报率	准确率	误报率	准确率	误报率
节点1	99.59	0.38	99.76	0.44	99.96	7.95	98.45	0.51	99.13	0.48	99.92	0.37
节点2	99.70	0.44	99.78	0.44	99.96	7.95	98.83	0.55	99.21	0.46	99.93	0.39
节点3	99.67	0.43	97.13	0.42	99.96	7.95	98.56	0.58	99.25	0.45	99.95	0.40
节点4	99.78	0.44	99.78	0.44	99.96	7.95	98.36	0.52	99.18	0.49	99.85	0.40
节点5	99.65	0.41	99.78	0.45	99.96	7.95	98.78	0.55	99.24	0.44	99.34	0.39
节点6	99.74	0.48	97.13	0.42	99.96	7.95	97.89	0.53	99.31	0.47	99.92	0.41
平均	99.69	0.43	98.89	0.44	99.96	7.95	98.48	0.54	99.22	0.47	99.82	0.39

表1显示,HDE-LSSVM 在6个节点的网络测试数据上有着比较稳定的检测准确率和误报率.虽然,HDE-LSSVM 并没有在每个节点相对其他方法取得最高准确率,但是总体的平均准确率次于投票方法.稳定的高准确率说明,HDE 求得的最优 LSSVM 模型可以达到稳定且准确分类的水平,提高了 LSSVM 的鲁棒性.值得一提的是,HDE-LSSVM 的误报率相对较低,这说明了 HDE 比 PSO 达到了更高精度的优化效果.

除此以外,还进行了标准差分演化算法(DE)和自适应差分演化算法(SaDE)优化 LSSVM 的仿真实验.实验结果表明,DE-LSSVM 和 SaDE-LSSVM 在检测准确率和误报率上都劣于 PSO-LSSVM 方法,这一现象也说明了 Dastanpour 等的观点:网络入侵检测源数据的多样性影响了演化算法优化 SVM 的精度.笔者提出的 HDE-LSSVM 算法则弥补了这一缺陷,显著提高了网络入侵检测的效率.

4 结论

笔者从优化最小二乘支持向量机的关键参数入手,用差分演化算法取得了更高、更稳定的优化效率;并针对网络数据的多样性,研究设计了多种扰动向量丰富了差分演化算法的扰动向量池,实现了 LSSVM 的自适应参数调优.实验结果表明,多扰动向量的策略大大提高了优化性能,并且使得 LSSVM 在网络入侵检测上有更稳定的平均性能.

参考文献:

[1] HU W M, GAO J, WANG Y G, et al. Online adaboost-based parameterized methods for dynamic distributed network intrusion detection[J]. IEEE transactions on cybernetics, 2014, 44(1): 66 – 82.

[2] 刘羿. 蝙蝠算法优化神经网络的网络入侵检测[J]. 计算机仿真, 2015, 32(2): 311 – 314.

[3] 李振刚, 甘泉. 改进蚁群算法优化 SVM 参数的网络入侵检测模型研究[J]. 重庆邮电大学学报(自然科学版), 2014, 26(6): 785 – 789.

[4] 王亚, 熊焰, 龚旭东, 等. 基于混沌 PSO 算法优化 RBF 网络入侵检测模型[J]. 计算机工程与应用, 2013, 49(10): 84 – 87.

[5] DASTANPOUR A, IBRAHIM S, MASHINCHI R, et al. Comparison of genetic algorithm optimization on artificial neural network and support vector machine in intrusion detection system[J]. Open systems, 2014, 77(10): 72 – 77.

[6] 马琰, 闫兵. 基于混沌差分优化算法的网络入侵检测系统[J]. 科学技术与工程, 2013, 13(36): 10967 – 10970.

[7] 边根庆, 赵宏, 张维琪, 等. 基于免疫克隆与差分进化的入侵检测方法[J]. 微电子学与计算机, 2012, 29(5): 124 – 128.

[8] SAILAGA M, KUMAR R K, MURTY P S R. Intrusion detection model based on differential evolution[J]. International journal of computer applications, 2011, 36(6): 10 – 13.

[9] STORN R, PRICE K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of global optimization, 1997, 11(4): 341 – 359.

[10] BREST J, ŽUMER V, MAUCEC M. Self-adaptive differential evolution algorithm in constrained real-parameter optimization[J]. IEEE congress on evolutionary computation, 2006, 98: 215 – 222.

[11] DENG W, YANG X, ZOU L, et al. An improved self-adaptive differential evolution algorithm and application[J]. Chemometrics and intelligent laboratory systems, 2013, 128(15): 66 – 76.

[12] YU Chengchi, CHEN J, HUANG Q, et al. A new hybrid differential evolution algorithm with simulated annealing and adaptive Gaussian immune[C]//2012 8th IEEE International Conference on Natural Computation (ICNC). Chongqin, China: IEEE, 2012: 600 – 607.

(下转第49页)

[5] 郭联哲,谭忠富,李晓军. 基于用户响应下的分时电价优化设计模型与方法[J]. 电网技术, 2006,30(5):25 – 28.
[6] 路郑. 国网大刀阔斧节能减排[N/OL]. 中国能源报,2015 – 06 – 22(18).
[7] 胡福年,汤玉东,邹云. 需求侧实行峰谷分时电价策略的影响分析[J]. 电工技术学报,2007,22(4):168 – 174.
[8] 杨桂元,郑亚豪. 多目标决策问题及其求解方法研究[J]. 数学的实践与认识,2012,42(2):108 – 115.
[9] 郜璘,周国祥,石雷. 峰谷分时电价决策的优化模型[J]. 统计与决策,2010(3):51 – 53.
[10] 陈缨,殷善锋. 基于用户需求弹性的峰谷分时电价决策模型[J]. 特区经济,2013(10):188 – 191.
[11] 陈沧杨,胡博,谢开贵,等. 计入电力系统可靠性与购电风险的峰谷分时电价模型[J]. 电网技术,2014,38(8):2141 – 2148.

Time-of-use Price Optimization Model Considering Line Loss

ZHAO Guosheng, ZHAN Tianle, LI Bo

(School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: The current TOU price optimization models failed to consider the significance of reducing line loss. Through the relationship between line loss and load fluctuation, the targets of the proposed new TOU price optimization model, which was based on price elasticity matrix of demand, were to minimize peak-valley difference and line loss. The optimization model, which was a non-linear multiple objects optimal model, was solved by ideal point method. It was proved by example that the new TOU price optimization model could avoid the peak load and reduce the line loss.

Key words: time-of-use price; line loss; price elasticity matrix of demand; multi-objective optimization model; ideal point method

(上接第 32 页)

Network Intrusion Detection Algorithm Based on Hybrid Differential Evolution Algorithm

WANG Yaoguang[1], Chen Weiquan[1], WU Zhenbang[1], QIN yong[2], HUANG Han[3]

(1. Guangdong Dongguan Quality Supervision Testing Center, Dongguan 523000, China; 2. School of Computer Science and Network Security, Dongguan University of Technology, Dongguan 523000, China; 3. School of Software Engineering, South China University of Technology, Guangzhou 510006, China)

Abstract: Intrusion detection algorithm based on machine learning method is one of research hotspot in the field of network equipment testing. In the face of the real-world application requirement, machine learning methods should be further optimized to achieve accurate and stable detection effect. The study optimize steadily several key parameters of least squares support vector machine (SVM) by designing a hybrid differential evolution algorithm with disturbance vector and improved the intrusion detection accuracy and stability of least squares support vector machine (SVM) algorithm by means of adaptive parameter tuning. The experimental results in KDD Cup 09 test set showed that, the proposed network intrusion detection algorithm based on hybrid differential evolution algorithm had better performance on average than many similar algorithm at present.

Key words: network intrusion detection; stability test; hybrid differential evolution; least squares support vector machine